10/785,990 MV-892

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 17/30 — A1

(11) International Publication Number: **WO 99/08209**

(43) International Publication Date: 18 February 1999 (18.02.99)

(21) International Application Number: PCT/US98/15862

(22) International Filing Date: 30 July 1998 (30.07.98)

(30) Priority Data:
08/907,705 8 August 1997 (08.08.97) US

(71) Applicant: BELL COMMUNICATIONS RESEARCH, INC. [US/US]; 445 South Street, Morristown, NJ 07960-6438 (US).

(72) Inventors: COCHINWALA, Munir; 14 Bullion Road, Basking Ridge, NJ 07920 (US). LEE, Kuo-Chu; 30 Revere Court, Princeton Junction, NJ 08850 (US). YU, Min, Tae; 3 Governor Drive, Basking Ridge, NJ 07920 (US).

(74) Agents: GIORDANO, Joseph et al.; International Coordinator, Rm. 1G112R, 445 South Street, Morristown, NJ 07960-6438 (US).

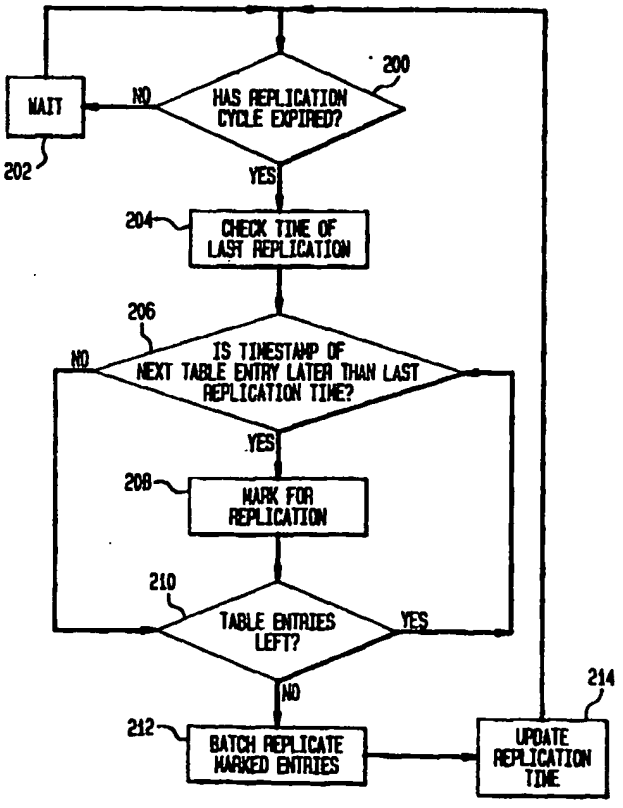
(81) Designated States: CA, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
With amended claims and statement.

(54) Title: HIGH PERFORMANCE ASYNCHRONOUS DATA REPLICATION

(57) Abstract

A table configuration and process (figure 1) enables efficient replication of table data (105). Time-stamp information (108) is included in a table (106) and each modification or update to a table entry is provided with a new time-stamp. A table replication (105) process checks the time-stamp of each table entry (106) periodically to identify those table entries which have been modified since a previous replication of the table. After marking those table entries which still require replication, the process writes only those identified table entries to a replication table. The replication cycle (200) may be selected at any effective time interval depending on the application.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

HIGH PERFORMANCE ASYNCHRONOUS DATA REPLICATION

Background of the Invention

The present invention relates to the field of databases, and more particularly to asynchronous data replication technology for highly available systems.

In various processing environments, data must be replicated to alternative storage databases at various times to ensure that the data is not lost. Processors in many stringent environments, such as telephone networks, have very specific data integrity requirements, which significantly increase the replication burden.

For asynchronous data replication, data is replicated periodically from one database to another. If one database fails, data is available from the other database. A replication interval is defined to control how frequently replication is performed.

Conventional replication approaches use a log to track data. For example, for each database operation, such as insert, update, or delete on a table to be replicated, a log is created, updated, or deleted to keep track of the operation. This approach results in the overhead of database operations due to the additional manipulation of the log. In addition, the size of the log is proportional to the number of operations. For instance, if there are several updates on a record, there will be the same number of logs even though only the last update may need to be replicated. This approach results in significant replication overhead. Thus, conventional asynchronous data replication techniques are limited by both operations and replication overhead.

Accordingly, it is an object of the present invention to overcome the disadvantages and drawbacks of conventional data replication systems and techniques.

Description of the Invention

To achieve these and other advantages and in accordance with the purposes of the invention, as embodied and broadly described, the invention includes a method of replicating a table, executed by a processor, comprising the steps of marking each entry to a table with a time-stamp corresponding to a time the entry is made to the table, checking the time-stamp of each entry in the table at a predetermined replication interval, identifying each table entry having a time-stamp after a known replication time, and writing only each identified table entry to a replication table.

The invention further includes a table replication system comprising a table including a time-stamp entry for each item entered in the table, a device for marking each entry in the table with a time-stamp corresponding to a time the entry is made to the table, a device for periodically identifying each table entry having a time-stamp after a predetermined time, and a device for writing each identified table entry to a replication table.

It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory and are intended to provide further explanation of the invention as claimed.

Brief Description of the Drawings

The accompanying drawings, which are incorporated in and constitute a part of the specification, illustrate presently preferred implementations of the invention and, together with the general description given above and the detailed description of the preferred implementations given below, serve to explain the principles of the invention.

In the drawings:

Fig. 1 is a block diagram of a computer and a table to be replicated in accordance with one embodiment of the present invention; and

Fig. 2 is a table replication process in accordance with one embodiment of the present invention.

Best Mode for Carrying Out the Invention

Reference will now be made in detail to the construction and operation of preferred implementations of the present invention which are illustrated in the accompanying drawings.

The following description of the preferred implementations of the present invention is only exemplary of the invention. The present invention is not limited to these implementations, but may be realized by other implementations.

Fig. 1 is a block diagram of a processing environment requiring a table to be replicated in accordance with one embodiment of the present invention. Processing environment 100 may correspond to any computer, processor, or network component that processes information, builds tables, and replicates those tables to a database, or any other form of data replication environment. Such a processing environment may exist, for example, in a reliable message

passing system of the telephone network, including, e.g. communications between a service management system and other network elements for "800 services."

As shown in Fig. 1, processing environment 100 includes a processor 102, a replication database 104, and a table 106. Replication database 104 corresponds to any suitable storage configuration for storing a replication table 105 corresponding to table 106. In accordance with conventional techniques, processor 102 builds a table 106 in processing environment 100 by adding, updating, and deleting data as necessary.

In accordance with the present invention, table 106 includes an additional "time-stamp" column 108. Each time processor 102 writes an entry to table 106 or performs an operation affecting a table entry in table 106. Processor 102 time-stamps that operation by inserting the time of the operation under the time-stamp column 108.

Fig. 2 provides a processing flow diagram of a replication process in accordance with one embodiment of the present invention.

Initially, processor 102 determines whether a replication cycle has expired (step 200). If not, it waits for a predetermined period (step 202) and returns to the initial step 200. The replication cycle preferably corresponds to a preselected time period for writing to the replication table in replication database 104. In a preferred embodiment, this replication cycle is chosen to be ten seconds. However, the replication cycle is configurable and can be selected to be any predetermined time period depending on the application.

If processor 102 determines that the replication cycle has expired in step 200, it next checks the time of the last replication (step 204). Against that information, processor 102 determines whether the time-stamp of the first table entry is later than the last replication time (step 206). If yes, processor 102 marks that table entry for replication (step 208). If the time-stamp of the table entry being checked is not later than the last replication time, processor 102 determines whether any table entries are left (step 210). Likewise, after processor 102 marks a particular table entry for replication, it checks for any table entries left (step 210). If a table entry still exists, processor 102 returns to step 206 and checks through each existing table entry. Once all table entries have been returned, processor 102 replicates each marked table entry by copying those marked table entries to the replication table 105 in replication database 104 (step 212). Processor 102 then updates the stored replication time (step 214) and returns to the beginning of the process.

In this manner, the present invention need not replicate an entire table of data. Instead, the process replicates only table entries that have been modified in some way since a last replication cycle.

While there has been illustrated and described what are at present considered to be preferred embodiments and methods of the present invention, it will be understood by those skilled in the art that various changes and modifications may be made, and equivalents may be substituted for elements thereof without departing from the true scope of the invention.

In addition, many modifications may be made to adapt a particular element, technique or implementation to the teachings of the present invention without departing from the central scope of the invention. Therefore, it is intended that this invention not be limited to the particular embodiments and methods disclosed herein, but that the invention include all embodiments falling within the scope of the appended claims.

Claims

1. A method of replicating a table, executed by a processor, comprising the steps of:

marking each entry to a table with a time-stamp corresponding to a time the entry is made or changed;

checking the time-stamp of each entry in said table at a predetermined replication interval;

identifying each table entry having a time-stamp after a known replication time; and

writing only each identified table entry to a replication table.

2. A method according to claim 1, further comprising the step of:

updating said replication time to correspond substantially to the time of said writing step.

3. A method of replicating a table, comprising the steps of:

marking each entry to a table with a time-stamp corresponding to a time the entry is made or changed;

periodically identifying each table entry having a time-stamp later than a predetermined time; and

writing each identified table entry to a replication table at substantially the same time.

4. A table replication system, comprising:

a table including a time-stamp entry for each item entered in said table;

means for marking each entry in said table with a time-stamp corresponding to a time the entry is made or changed;

means for periodically identifying each entry having a time-stamp after a predetermined time; and

means for writing each identified table entry to a replication table.

AMENDED CLAIMS

[received by the International Bureau on 15 December 1998 (15.12.98);
original claims 1-4 replaced by new claims 1-4 (1 page)]

1. In a data processing system having a main table for storing data and wherein the data in the main table is replicated in a replication table, a method for reducing the replication load comprising the steps of

marking each entry made in the main table with a time-stamp corresponding to the time the entry is made or changed,

checking the time-stamp of each entry in the main table at predetermined replication intervals,

identifying each table entry in the main table having a time-stamp after a prior replication time, and

writing only each such identified table entry in the replication table.

2. The method according to claim 1, further comprising updating said prior replication time to correspond to the time of said writing step.

3. In a data processing system having a main table for storing data and wherein the data in the main table is replicated in a replication table, a method for reducing the replication load comprising the steps of

marking each entry in the main table with a time-stamp corresponding to the time the entry is made or changed,

periodically identifying each main table entry having a time-stamp later than a predetermined time, and

writing only each such identified table entry in the replication table at substantially the same time.

4. In a data processing system, the combination comprising

a main table,

a replication table,

means for marking each data entry in said main table with a time-stamp corresponding to the time said each entry is made or changed,

means for periodically identifying each data entry in said main table having a time-stamp after a predetermined time, and

means for writing only each such identified data entry from said main table to said replication table.

Statement Under Article 19

The amendments being made to claims 1 through 4 as originally filed clarify the language of the claims and improve their form. The amendments have no impact on the description and drawing of the application.

FIG. 1

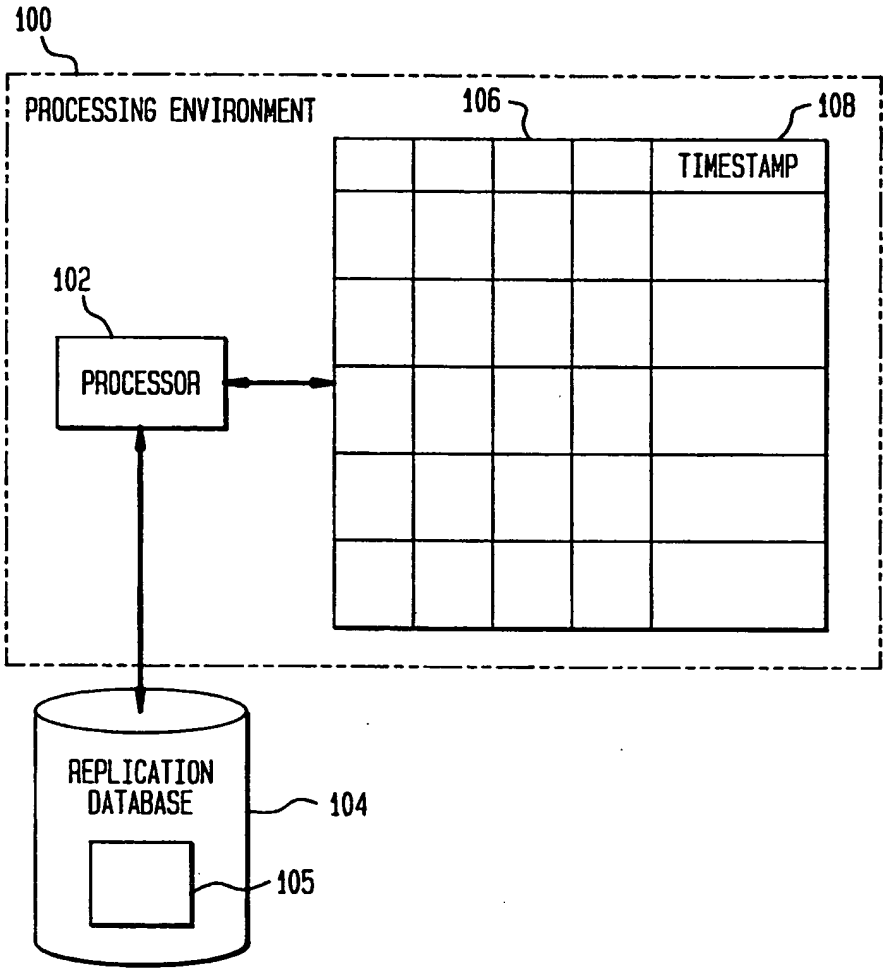
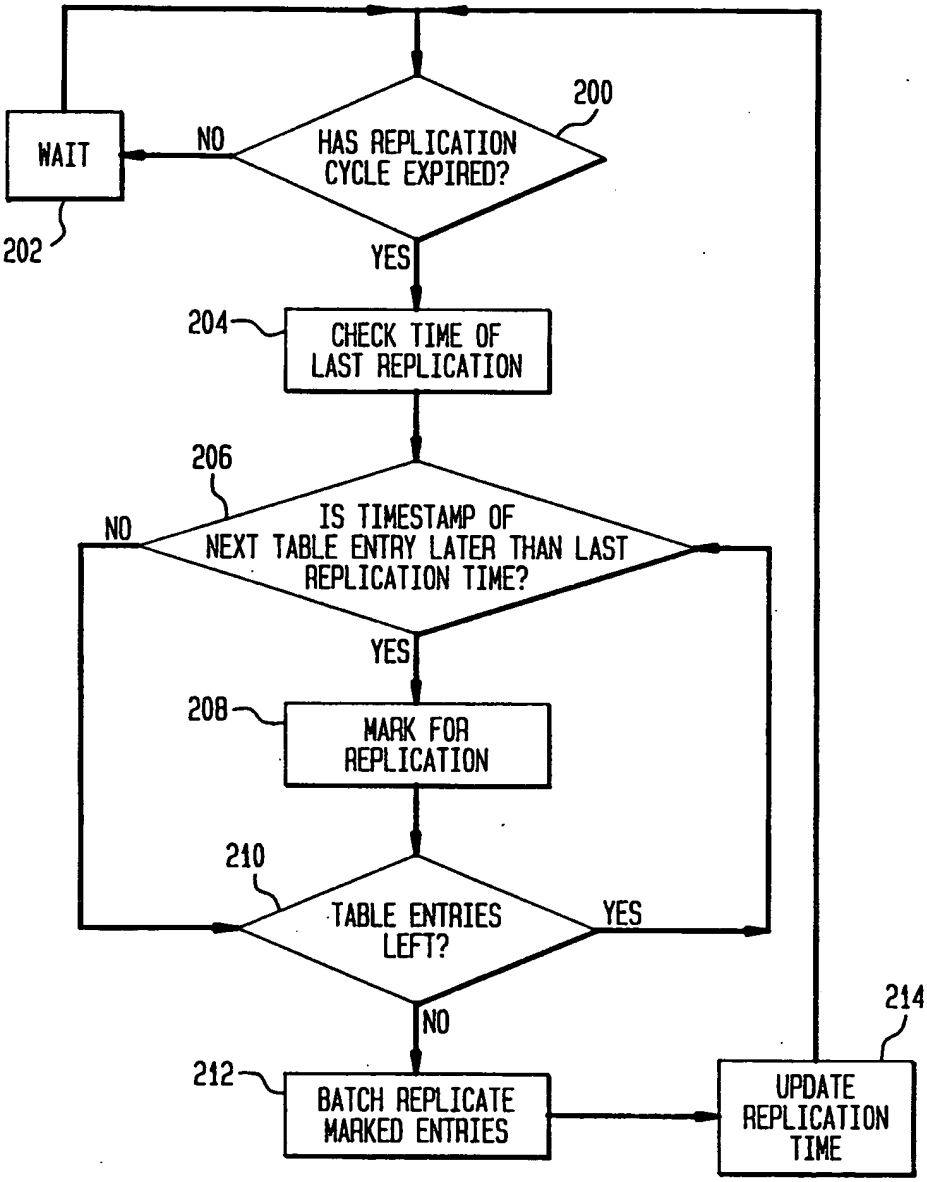

100
PROCESSING ENVIRONMENT
106
108
TIMESTAMP
102
PROCESSOR
REPLICATION DATABASE
104
105

FIG. 2


200
HAS REPLICATION CYCLE EXPIRED?
NO
WAIT
202
YES
204
CHECK TIME OF LAST REPLICATION
206
IS TIMESTAMP OF NEXT TABLE ENTRY LATER THAN LAST REPLICATION TIME?
NO
YES
208
MARK FOR REPLICATION
210
TABLE ENTRIES LEFT?
YES
NO
212
BATCH REPLICATE MARKED ENTRIES
214
UPDATE REPLICATION TIME

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/15862

A. CLASSIFICATION OF SUBJECT MATTER
IPC(6) :G06F 17/30
US CL :707/1-5,8,10,100-103,200-203
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 707/1,2,3,4,5,8,10,100,101,102,103,,201,202,203; 395/673,674,182.13,200.32,183.21; 364/282.1,282.2,281.8

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,603,024 A (GOLDRING) 11 FEBRUARY 1997, SEE COL. 3, LINES 8-47; COL. 4, LINES 64-67; COL. 5, LINES 1-20, 48-67; COL. 6, LINES 1-27; COL. 7, LINES 38-56.	1-4
A,P	US 5,778,225 A (SUPERNAW-ISSEN ET AL.) 07 JULY 1998	1-4
A,P	US 5,758,355 A (BUCHANAN) 26 MAY 1998, SEE ENTIRE DOCUMENT.	1-4
A,P	US 5,737,600 A (GEINER ET AL.) 07 APRIL 1998, SEE ENTIRE DOCUMENT.	1-4
A	US 5,544,353 A (FORMAN ET AL.) 06 AUGUST 1996, SEE ENTIRE DOCUMENT.	1-4

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search	Date of mailing of the international search report
28 SEPTEMBER 1998	29 OCT 1998
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer SRIRAMA CHANNAVAJJALA Telephone No. (703) 308-8538

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/15862

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	SIDELL ET AL., Data Replication in Maripose, IEEE. February 1996, pages 485-494.	1-4
A	PARK ET AL., A Replica Control Method for Improving Availability for read-only transactions, IEEE, April 1997, pages 104-112.	1-4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/15862

B. FIELDS SEARCHED
Electronic data bases consulted (Name of data base and where practicable terms used):

APS, DIALOG
search terms: replication databases, replication data, asynchronous datas, telephone networks, tables, time stamps, replicat or copy, log data, distributed databases, RDBMS or relational database management system, checking time stamp, predetermined intervals